基于改进多层次模糊关联规则的定量数据挖掘算法 *

张定祥 [1]，张跃进 [2]

(1. 贵州商学院 计算机与信息工程学院, 贵阳 550014; 2. 华东交通大学 信息工程学院, 南昌 330013)

摘　要： 针对单一层次结构实现规则提取，具有规则提取准确性不高，算法运行时间长，难以满足用户使用需求的问题，提出一种基于改进多层次模糊关联规则的定量数据挖掘算法。采用高频项目集合，通过不断深化迭代的方法形成自顶向下的挖掘过程，整合模糊集合理论、数据挖掘算法以及多层次分类技术，从事务数据集中寻找模糊关联规则，挖掘出储存在多层次结构事务数据库中定量值信息的隐含知识，实现用户的定制化信息挖掘需求。实验结果表明，提出的数据挖掘算法在挖掘精度和运算时间方面相较于其他算法具有突出优势，可为多层次关联规则提取方法的实际应用带来突破性进展。

关键词： 模糊集合；用户定制化；多层次结构；柔性边界；隶属度函数

中图分类号： TP311　　doi: 10.3969/j.issn.1001-3695.2018.06.0405

Quantitative data mining algorithm based on improved multi-level fuzzy association rules

Zhang Dingxiang[1], Zhang Yuejin[2]

(1. *College of Computer & Information Engineering Guizhou University of Commerce*, *Guiyang* 550014, *China*; 2. *School of Information Engineering*, *East China Jiaotong University*, *Nanchang* 330013, *China*)

Abstract: In order to extract rules from a single hierarchy, the accuracy of the rule extraction is not high, the algorithm runs long, and it is difficult to meet the needs of the users, this paper proposes a quantitative data mining algorithm based on the improved multilevel fuzzy association rules, adopt the high frequency project set, form the continuous deepening of the iterative method. In the top down mining process, fuzzy set theory, this method integrates data mining algorithm and multi-level classification technology to find fuzzy association rules from the transaction data set, excavates the hidden knowledge of quantitative value information in the multi-layer structured transaction database, and realizes the user's customized information mining needs. The experimental results show that the quantitative data mining algorithm based on the improved multilevel fuzzy association rules has a prominent advantage over other algorithms in mining precision and operation time. It can bring about breakthrough in the practical application of multilevel association rule extraction.

Key words: fuzzy set; user-defined; multi level structure; flexible border; belonging function

0　引言

近年来，随着数据科学领域理论体系和算法的日益完善，基于数据算法的科学研究理论正逐渐成为学术界和工业界的关注焦点[1]。其中，数据挖掘理论作为数据关系信息提取的重要方法而成为数据科学领域研究的重点。根据所挖掘数据信息的不同，可以将数据挖掘方法进一步细分为关联挖掘、分类挖掘、聚类挖掘以及序列挖掘等[2]。关联挖掘是数据挖掘的重要类型，该方法主要用于确定事务数据库中不同项目之间的相关性。关联挖掘方法已被广泛应用于市场规划和营销策略制定等领域，并取得了较好的应用效果[3]。例如，超市的管理人员可以使用关联挖掘有效预测人们更倾向于一起购买的商品组合，类似于“购买纸尿裤的顾客通常也会购买啤酒”之类的关联规则就可以被挖掘出来。基于这些关联规则，超市的管理人员可以将啤酒和纸尿裤摆放在超市相近的位置来诱导顾客同时购买。可见，关联规则的定量数据挖掘研究意义重大。

纵观近年来关联挖掘算法学术研究成果，大多基于 Aprior 算法通过逐步产生并测试候选项目集合实现[4,5]。然而这一过程通常需要对数据库进行遍历扫描计算，数据计算成本较高。随着关联挖掘数据样本容量呈现指数级增长趋势，Aprior 算法所

收稿日期: 2018-06-26; **修回日期**: 2018-08-09　**基金项目**: 国家自然科学基金资助项目（61164013）; 贵州省软科学研究计划项目（黔科合体 R 字［2014］LKS2007）; 贵州省教育厅基金项目（黔教社发［2010］339）; 贵州省普通高等学校智能物联网工程研究中心建设项目（黔教合 KY 字［2016］016）; 贵州省教育厅项目（黔教合 KY 字［2017］022）

作者简介: 张定祥（1969-），男（苗族），贵州松桃人，高级工程师，副教授，主要研究方向为数据挖掘（zhangdingxianggz@126.com）; 张跃进（1978-），男，湖北钟祥人，副教授，博士，主要研究方向为数据挖掘、计算机应用.

耗费的高昂时间成本已成为关联挖掘研究领域亟待解决的关键问题[6]。因此，文献[7]提出关联规则需要满足自信度和支持度两种用户特定的约束度，以降低数据挖掘的计算时间。其中，支持度定义为事务集中满足条件的事务所占的比例，而自信度定义为满足条件的事务支持度与事务集支持度的比值。

此外，目前绝大多数关联规则算法研究成果都仅仅着眼于单一概念层次挖掘，对于多概念层次挖掘较少涉及，如文献[8]、文献[9]的算法。文献[10]提出模糊挖掘算法，应用多层次关联挖掘，从关联挖掘方法实际应用需求角度出发，为用户提供更多有价值的信息。但是在提供应对多余规则的解决方案时，迭代算法复杂，需要耗费较高计算资源。

针对上述研究现状，提出一种基于改进多层次模糊关联规则的挖掘算法，可用于提取定量数据中的隐含信息。该方法采用高频项目集合，通过不断深化迭代的方法形成自顶向下的挖掘过程。算法整合了模糊集合理论、数据挖掘算法以及多层次分类技术，着眼于从事务数据集中寻找模糊关联规则。实验结合具体算例验证该算法的优越性，对于用户来说，该方法挖掘得到的规则更具有逻辑性，且更符合人类思维认知。

1　提出的改进多层次模糊关联规则挖掘算法

为了挖掘多层次关联规则，需要对项目进行分类或者对概念的层次结构进行有效定义[11]。其中概念的层次结构可以从一个有向无环图（directed acyclic graph, DAG）[12]复制得到。概念的层次结构代表项目的通路与需求之间的关系，并能将它们在不同的抽象层级上分类。这些概念层次具有可用性，或者可以通过某一领域的专家应用得到。例如一个用户通常不仅关心电脑与打印机之间的关联，而更希望得到台式电脑的价格与激光打印机的价格之间的关联。此外，模糊理论[13]对于多层次关联挖掘方法的研究具有一定借鉴意义，这一理论提出通过引入渐进成员关系来表征语言术语的模糊边界[14]。

因此，为实现定量数据集中多层次关联规则的有效挖掘，将基于分类学理论成果[15]，提出一种改进的模糊挖掘算法。这一算法综合利用数据挖掘方法、多层次分类理论以及隶属函数定义，可用于在给定的事务数据集中挖掘模糊关联规则。

1.1　改进多层次关联规则

在多概念层次上挖掘关联规则可能会获得更具普适性和可用性的规则。具体项目的分类在实际应用场景中通常是预先定义，并且能够用结构树进行表示的。结构树的终端节点代表事务中出现的实际项目；内部节点表示低层次节点所形成的概念或类别。

在图 1 中，根节点位于第 0 级，表示分类的内部节点（如饮料）位于第 1 级，表示口味的内部节点（如柠檬味）位于第 2 级，而表示品牌（如可口可乐）的终端节点位于第 3 级。最终在算法流程中只有终端节点出现在事务中。在预定义的分类中，根据在结构树中所处的位置，各节点首先被编码为数字和符号“*”的组合。例如，图 1 中的内部节点“果汁”被编码为 1**，内部节点“草莓味”被编码为 11*，终端节点“汇源”被编码为 111。

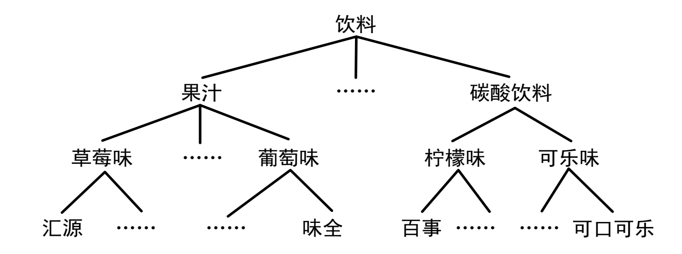


图 1　基于饮料分类的结构树编码示意图

1.2　多层次模糊关联挖掘算法建立

多层次模糊关联挖掘算法的算法建立步骤如下所示：

利用数据集和符号“*”组成的符号序列，对预定义的分组进行编码，该序列可以根据式(1)得到

$$D = S \times 10 + j \tag{1}$$

其中：j 为节点在当前层次的位置序号（节点的位置序号是从 1 开始的连续整数，每个节点按照从左到右的顺序依次编码）；D 为当前层次中该节点的编码；S 为当前层次该节点的父节点编码。

依据式(1)提供的编码法则，可以为一个结构层次中的任意节点编码。为了便于展示编码流程的具体操作，图 2 以一个典型的四层结构为例展示了每个节点的编码。在编码完成之后，事务数据库中每个项目都会替换为它对应的编码。

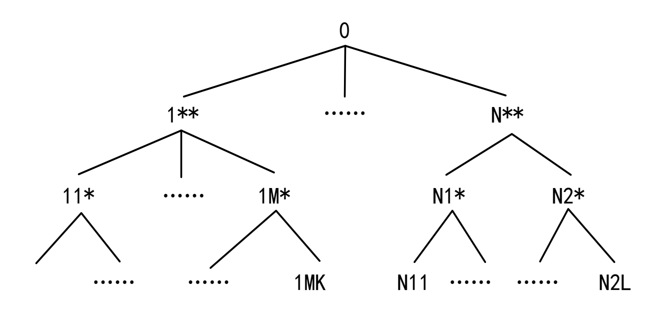


图 2　一个典型四层结构编码案例图

之后，令 $k=1.r=1$，其中：$1 \le k \le x$ 是当前的层次序号；x 是规定分类结构的层次数量；r 表示存储在当前的频繁项目集合中的项目数量。

对于每一个事务数据 D_i，i 表示事务序号，其上限为数据库中事务数据的总量，将前 f 位相同的项目加在一起，计算它们的支持度，并将支持度小于当前层次最小支持度 ʞ 的小组移除。

对于不同的数据项目，分别预设不同的隶属函数来表征各类项目的差异性。对于每一个不同的数据项目，都具有其独特的属性以及隶属度函数，之后将每个事务数据 D_i 的分组值转换为模糊集合，这种转换可以通过特定的隶属函数映射得到。具体转换的公式如式(2)所示。

$$\sum_{c=1}^{h} (f_{abc}^{s} / R_{bh}^{s}) \tag{2}$$

对事务数据集中的所有事务按照式(2)的方法进行模糊集合的组合及划分。根据式（3）计算每个模糊区域 Z_{al}^{s} 在十五数据中的值，其中 Sum_{al}^{s} 是所有 Z_{al}^{s} 的和。

$$Sum_{al}^{s}=Z_{1bl}^{s}+Z_{2bl}^{s}+Z_{3bl}^{s}+......+Z_{nbl}^{s} \tag{3}$$

然后，根据式（4）指定 $setsum_b^s$。其中 $setsum_b^s$ 是 Sum_{al}^s 的最大值（$1 \le l \le h_j^k$）。

$$setsum_b^s = Maximum[_{l=1}^{p}(sum_{bl}^{s})] \tag{4}$$

令 $setsum_b^s$ 是项目 Z_{al}^s 中具有 $setsum_b^s$ 的区域。如果区域 $\max R_j^k$ 的值 $setsum_b^s$ 在当前层次大于或等于最小支持度（K），那么就将 $setsum_b^s$ 放置于频繁 1-项目集合。

针对不同的层次序号值，执行下面不同的过程：

a)如果在第 2 个层次结构中产生了候选集合 C_2^k，其中 C_2^k 表示第 k 层具有多个候选项目的集合，则说明该算法可以继续应用，这些项目都是从各个层次中通过模糊层次交叉方法得到的频繁项目。例如，层次 2 上的候选 2-项目集合并不仅仅局限于层次 2 上的频繁项目对，层次 2 上的频繁项目也可能与层次 1 上的频繁项目组合形成层次 2 上的候选 2-项目集合。但是根据分类算法基本理论可知，每一个候选项目集合中的 2-项目集合都必须包含至少一个 L_1^k 中的项目，并且下一个项目不是该项目在分类学上的祖先。所有可能的 2-项目集合都被收集在 C_2^k 中。得到这一集合后则开始执行步骤 b)。

b)如果层次结构序号>2，需要通过软件方法产生候选集合 T_r^k，T_r^k 是层次 k 上由 T_r^{k-1} 产生的具有多个项目的候选项目集合，其产生方法与 apriori 算法产生候选项目集合的方法类似。

对于任意在 T_r^k 中通过筛选获得的候选 r-项目集合：

$$M=(M_1,M_2,...,M_r)$$

a)计算该集合中每一个事务数据下的模糊值，该计算需要通过式(5)中的算法进行。

$$P_{it}=\min imum(P_{is_1},P_{is_2},...,P_{is_r}) \tag{5}$$

b)令 Sum_{al}^s 是 f_{is} 的和，$1 \le i \le n$，即

$$Sum_{tl}^{s}=f_{1tl}^{s}+f_{2tl}^{s}+f_{3tl}^{s}+......+f_{ntl}^{s} \tag{6}$$

c)如果 $Count_s$ 在当前层次不小于最小支持度 K，就将 $Count_s$ 插入 T_r^k。

d)选择所有满足自信度不小于预定义的自信度阈值 T 的规则，其中 T 是预定义的最小自信度。

2 多层次模糊关联挖掘算例分析

为了具体阐述该算法的应用流程和效果，结合具体算例对该算法进行实证性分析，在该算例中使用快消品零售超市中商品的销售作为事务。为简化验证过程，共随机选择七个事务，如表 1 所示。

表 1 快消品销售事务定义表

事务编号	项目
D_1	（婴幼儿护理产品，2）
	（沐浴露及洗发水，6）
	（防晒霜系列产品，3）
	（口腔护理品，4）
	（鞋及衣物护理产品，5）
D_2	（空气清新剂，6）
	（婴幼儿护理产品，5）
	（饮用水及饮料，3）
	（防晒霜系列产品，3）
	（巧克力食品，8）
	（鞋及衣物护理产品，4）
	（宠物护理产品及食物，6）
D_3	（巧克力食品，5）
	（饮用水及饮料，7）
	（婴幼儿护理产品，6）
	（沐浴露及洗发水，4）
D_4	（空气清新剂，1）
	（防晒霜系列产品，8）
	（巧克力食品，3）
D_5	（口腔护理品，8）
	（鞋及衣物护理产品，5）
D_6	（防晒霜系列产品，4）
	（巧克力食品，9）
D_7	（口腔护理品，9）
	（防晒霜系列产品，8）

使用预定义分类法，它们的分类如图 3 所示。

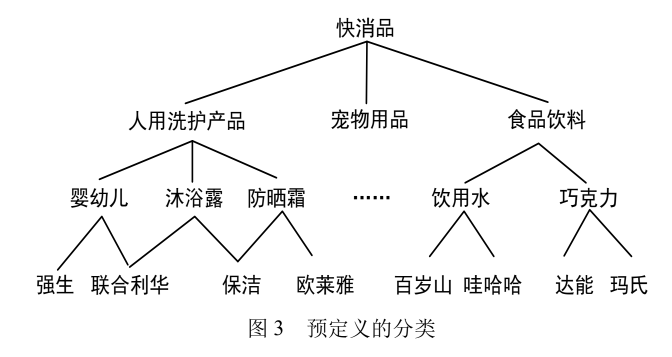


图 3 预定义的分类

如图 3 所示，将终端零售店所销售的快消品分为三类，分别为人用洗护产品、宠物产品和食品饮料类。每一类都可进一步细分为若干子分类，以确定快消品的细分行业和对应品牌。对于每一类的快消品，都有一个具有唯一性的隶属度函数，根据隶属度函数计算结果可以进一步把各项目划分为隶属度低、中、高三种模糊区域。

首先，将图 3 所示的快消品节点分类转换为其等价编码，其结果如表 2 所示。

表 2 例子的编码后事务数据

事务编号	项目
D_1	(111,2) (112,6) (211,3) (212,4)(311,5)
D_2	(111,6) (112,5) (212,3) (222,8) (322,4) (321,6)
D_3	(211,4) (221,7) (312,2) (322,8)
D_4	(112,10) (221,11) (313,6)
D_5	(112,7) (223,6)
D_6	(122,9) (142,22) (323,6) (333,9)
D_7	(111,9) (122,8)

令该层次结构的两个变量 μ 和 τ 的值均为 1。其中，μ 表示

当前项目所处的分类层次；τ 表示当前频繁项目集合的项目数量。

将数据库中所有 μ 相似的事务都合并为一个大类并将它们加和。例如，可以将项目（223,2）和（254,2）整合为（2**，4）。这一任务的结果如表 3 所示。

表 3　例子中的 1 级表示

事务编号	项目
D_1	(1**,8) (2**,7) (3**,5)
D_2	(1**,14) (2**,15)
D_3	(2**,11) (3**,10)
D_4	(1**,10) (2**,11) (3**,6)
D_5	(1**,7)(2**,6)
D_6	(1**,31)(3**,15)
D_7	(1**,17)

根据对应的隶属度函数，将所得到的组转换为模糊集的形式。以（1**，8）为例，根据图 3 中预定义的分类，这个组是属于人用洗护产品类的，需要使用前述的人用洗护产品隶属度函数。在这一隶属度函数中，计算结果为 6，对应着低区域的隶属度为 0.6，中区域隶属度为 0.9，高区域隶属度为 0.1。通过这种方式可以计算得到事务中的所有项目构成的等价模糊集。

在所有事务中计算每一个模糊区域值的和，得到各模糊区域隶属度之和，如表 4 所示。

表 4　各事务 1 级模糊区域隶属度之和计数表

项目	计数
（1**.低）	1.1
（1**.中）	1.5
（1**.高）	0.6
（2**.低）	1.2
（2**.中）	3.3
（2**.高）	1.8
（3**.低）	2.5
（3**.中）	0.6
（3**.高）	1.6

基于表 4 中归纳得到的各事务隶属度之和的计算结果，选择每组值最高的模糊区域，依次挑选出各组中计量值较高的模糊区域。在上一步完成之后，将各组中挑选出的模糊区域的隶属度分别与预定义的第 μ 层的最小支持度进行比较，并加入 P_1^1。例如，假定第一层的最小支持度为 1.3，从表 4 来看， 1**.中、2**.中、3**.低均大于或等于 1.3，这些频繁成员集合被放置于 P_1^1 中。候选项目集合 D_2^1 由 P_1^1 产生，由于 P_1^1 由 1**.中、2**.中和 3**.低三个成员组成， D_2^1 的成员如表 5 所示。

表 5　第 2 层的候选项目集合

项目集合
（1**.中，2**.中）
（1**.中，3**.低）
（2**.中，3**.低）

对 D_2^1 中每一个 2 成员项目集合执行下列步骤：

a） D_2^1 项目集合中的每一个 2 成员项目的模糊隶属度都根据所处事务每个项目的预定义隶属度函数进行计算，以项目集合 $\{2^{**}.中,3^{**}.低\}$ 为例。这一集合在事务 D_1 中的隶属度可以根据式(8)计算。

$$\min(1.1, 0.8) = 0.8 \tag{7}$$

利用该方法对所有事务进行模糊隶属度计算，得到的结果如表 6 所示。

表 6　模糊隶属度计算表

项目编号	2**.中	3**.低	Min（2**.中，3**.低）
D_1	1.1	0.8	0.8
D_2	0.8	1.3	0.8
D_3	0.7	0.4	0.4
D_4	0.7	0.9	0.7
D_5	0	0.5	0
D_6	0.5	0.8	0.5
D_7	1.2	0	0

b) 根据 A 部分的方法，可以计算 C_2^1 中每一个 2-成员集合的模糊隶属度的和。

c) 根据所得到的项目集合，只有（2**.中、3**.低）的结果大于预定义的第 1 层最小支持度 1.3，因此 C_2^1 集合中只有这一个成员。令 s=2，其中 s 表示当前项目集合中项目的数量。由于 C_2^1 只有一个 2-成员集合，无法在第 2 层上产生一个 3-成员集合。本文在 μ 中添加了一个单元，进入了步骤 b)。令层次 2 和层次 3 的 $\min supp = 2$，则这两层的频繁项目集合分别如表 7 和 8 所示。由于不存在第 4 级，因而可以直接执行下一步。

表 7　层次 2 的频繁项目集合

项目集合	计数
（2*.中）	3
（21*.中）	2
（31*.高）	2.2
（33*.中）	1.5
（32*.高）	2

表 8　层次 3 的频繁项目集合

项目集合	计数
（2**.低）	2.8
（21*.中）	1.8
（211.中，3**.低）	3.1
（2**.低, 3**.低）	1.3

基于前述步骤得到的频繁项目集合，可以开展模糊关联规则的挖掘：从各层级频繁项目集中按照下列规则检索所有可能的规则。需要注意的是，必须从包含最小二元项目的频繁项目集合中提取规则。具体规则集如下所示：

如果 2**=中则 3**=低；

如果 3**=低则 2**=中；

如果 3**=低则 21*=中；

如果 21*=中则 3**=低；

如果 211=中则 3**=低；

如果 3**=低，则 211=中。

为了获得符合用户指定条件的规则，必须计算每条规则对应的自信度

所得结果如表 9 所示。

表 9　所有规则的自信度

关联规则	自信度
如果 2*=中则 3**=低	1.2
如果 3**=低则 2*=中	1.1
如果 3**=低则 21*=中	0.8
如果 21*=中则 3**=低	0.6
如果 211=中则 3**=低	1.2
如果 3**=低，则 211=中	1.3

将表 9 中所有规则的自信度计算结果与预定义的最小自信度阈值进行比较，并保留自信度值大于预定义最小自信度阈值的规则作为最终规则挖掘结果。例如，如果最小自信度阈值设置为 0.9，最终的规则就是：

如果 2*=中则 3**=低；

如果 3**=低则 2*=中；

如果 211=中则 3**=低；

如果 3**=低，则 211=中。

3　实验对比与分析

文中提出了基于多层次模糊关联规则的数据挖掘算法，并通过具体算例详细阐述了算法应用流程，验证了其在实际应用中的可行性与有效性。为了进一步分析在大量数据信息工况下该算法性能的优越性，于 PC 机上通过 MATLAB 仿真平台利用所提出的算法对 1 000 张便利店购物票据信息进行分析。PC 配置为 i5 酷睿双核、8 GB 运行内存。

便利店中所有的商品可以分为 6 类，每一类都具有预定义的隶属度函数。基于购物票据上的信息和预定义的数据来挖掘这些项目之间的关联规则。预定义类别在第一层具有 6 个节点，表示测试中的项目名称；第二层有 12 个节点，分别表示细分种类或其他特定产品的不同类别信息；第三层有 45 个节点，分别表示这些产品的生产公司和厂家。

购物票据上的交易信息包含商品名称、型号、单价和商品的购买量。在每一笔交易中，不能多次包含同一个项目。

图 4 展示了 1 000 笔交易中，最小支持度为 3 的情况下挖掘到的规则数量与不同预定义最小自信度之间的关系。随着研究中交易数量的增加，所挖掘到的规则数量也会逐渐增加。这是因为随着交易数量的增加，频繁项目的数量也会增加，从而导致通过关联挖掘可以在最小支持度一定的情况下得到更多的规则。与此同时，根据结果可知，增加预定义的最小自信度的值，会使得所挖掘到的关联规则数量下降。

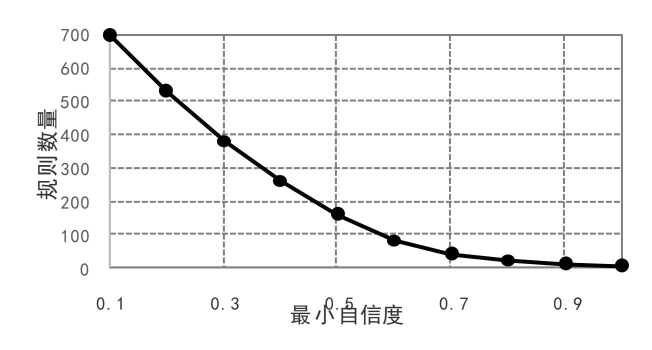


图 4　不同最小自信度下的关联规则数量

所提出多层次模糊关联挖掘算法的一个重要优点就是具有可以根据用户的需要，在不同的层次上挖掘关联规则的能力。即在提出的算法中，用户可以精确地指定需要挖掘哪一层次的规则，从而保证获得的结果能够最大限度地满足用户的使用需求。这是因为在所提出的算法中，不同层次的最小支持度是可以分别定义的，所以可通过提高某一层次最小支持度的值，使得程序在该层挖掘到的规则数量为 0。图 5 和 6 展示了 1 000 次交易中提出的算法与其他方法在结构层次 1 和 2 中挖掘到的规则数量与预定义最小支持度之间的关系。从图中可以看出，提出的算法相比于其他方法在不同的最小支持度下，能够更精准地获得规则数量，即能够挖掘出更小范围的符合要求的商品，算法挖掘精度更高，其中，最小自信度的值均取为 0.2。

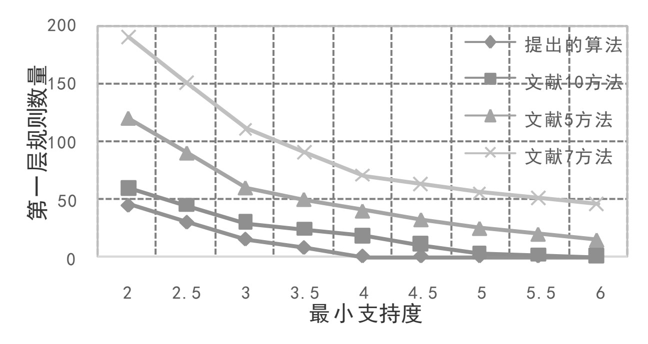


图 5　第 1 层中不同最小支持度对应的挖掘规则数量

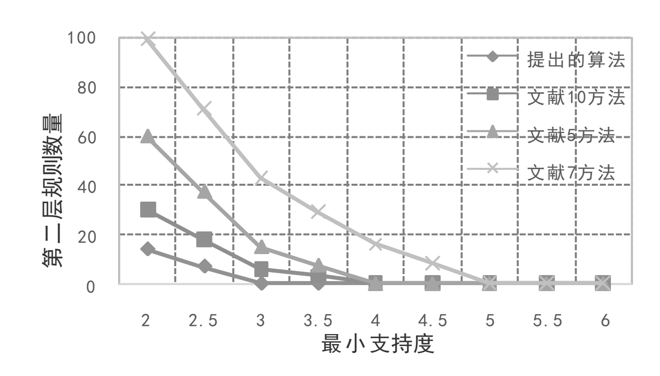


图 6　第 2 层中不同最小支持度对应的挖掘规则数量

根据关联挖掘算法应用需求调研可知，算法的运行时间和占用的计算资源是用户重点考量的因素之一。若算法运行时间过长，将在很大程度上失去对用户的吸引力。图 7 展示了所提出的算法与其他算法在处理不同交易数量时的运行时间性能对比。在文献[5,7]所提出的算法中，所有的分类标准都是由单值定义的，最小支持度和隶属度函数也都对应于所有的项目。为

了保证算法控制变量对比，挖掘最小自信度为 0~6 时的第一层和第二层的规则。图 7 展示了算法运行时间的比较结果，结果表明所提出的算法相比于文献[5,7]和[10]的方法，在不同最小支持度的情况下，运行时间均更短。这是因为提出的方法中运用改进的挖掘关联规则，能够更高效地获得具有普适性和可用性的规则，可见提出的算法不仅能按照用户的意愿挖掘不同层次的规则，而且还可以减少程序的运行时间，有利于用户满意度的极大提升。

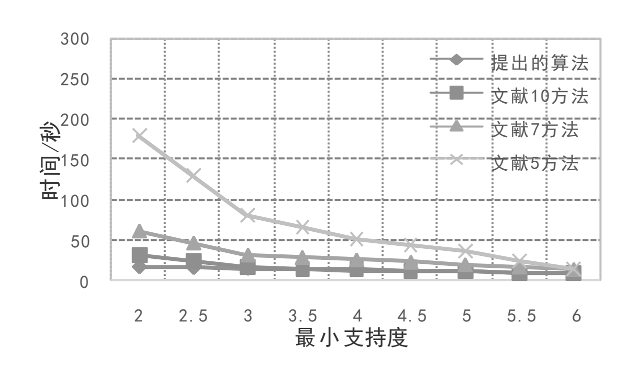


图 7　不同最小支持度下算法运行时间比较

4　结束语

在详细阐述现有关联规则挖掘算法的研究现状和主要挑战的基础上，综合利用模糊集理论、多层结构分类法以及数据挖掘理论，本文提出了一种基于多层次模糊关联规则的挖掘算法，可用于提取定量数据中的隐含信息。该方法采用高频项目集合，通过不断深化迭代的方法形成自顶向下的挖掘过程，具有根据用户的倾向挖掘不同层次的关联规则的能力，能够为不同的项目定义不同的隶属度函数，从而满足不同种类商品的定制化分析需求。

通过对快消商品终端店铺和便利店历史数据库的关联挖掘实验，证明了与相关研究成果相比，所提出基于多层次模糊关联规则算法挖掘精度更高、并且能够显著减少算法的计算时间，有利于用户满意度的提升。

参考文献：

[1] 顾玉萍，程龙生. 基于 MTS-AdaBoost 的不平衡数据分类研究 [J]. 计算机应用研究，2018，35 (2)：346-348. (Gu Yuping, Cheng Longsheng. Research on unbalanced data classification based on MTS-AdaBoost [J]. Application Research of Computers, 2018, 35 (2): 346-348.)

[2] 崔一辉，宋伟，王占兵，等. 一种基于格的隐私保护聚类数据挖掘方法 [J]. 软件学报，2017，28 (9)：2293-2308. (Cui Yihui, Song Wei, Wang Zhanbing, *et al.* A lattice-based clustering data mining method for privacy preservation [J]. Journal of Software, 2017, 28 (9): 2293-2308.)

[3] Pemajayantha V, Pemajayantha V, Mellor R, *et al.* Approaches of discriminant analysis for data mining and management [J]. Science, 2018, 200 (4349): 1481-1483.

[4] Wijayanti A W. Analisis hasil implementasi data mining menggunakan algoritma apriori pada apotek [J]. 2017, 3 (1): 60-69.

[5] 谢志明，王鹏. 一种基于 MapReduce 架构的并行矩阵 Apriori 算法 [J]. 计算机应用研究，2017，34 (2)：401-404. (Xie Zhiming, Wang Peng. A parallel matrix a priori algorithm based on MapReduce architecture [J]. Application Research of Computers, 2017, 34 (2): 401-404) .

[6] 安相华，于靖博，蔡卫国. 基于混合多属性决策和关联分析的模糊粗糙 FMEA 评估方法 [J]. 计算机集成制造系统，2016，22 (11)：2613-2621. (An Xianghua, Yu Jingbo, Cai Weiguo. Fuzzy rough FMEA evaluation method based on mixed multi attribute decision making and association analysis [J]. Computer Integrated Manufacturing System, 2016, 22 (11): 2613-2621.)

[7] 赵学健，孙知信，袁源. 基于预判筛选的高效关联规则挖掘算法 [J]. 电子与信息学报，2016，38 (7)：1654-1659. (Zhao Xuejian, Sun Zhixin, Yuan Yuan. An efficient association rule mining algorithm based on pre-selection [J]. Journal of Electronic and Information Science, 2016, 38 (7): 1654-1659.)

[8] García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining [J]. Knowledge-Based Systems, 2016, 23 (7): 98: 1-29.

[9] Ghaffarian S M, Shahriari H R. Software vulnerability analysis and discovery using machine-learning and data-mining techniques: a survey [J]. ACM Computing Surveys, 2017, 50 (4): 1-36.

[10] Yang J, Li J, Liu S. A new algorithm of stock data mining in Internet of multimedia things [J]. Journal of Supercomputing, 2017 (9): 1-16.

[11] 许栋浩，李宏伟，张铁映，等. 利用改进粒子群算法的关联规则挖掘 [J]. 测绘科学，2016，41 (2)：168-172. (Xu Donghao, Li Hongwei, Zhang Tieying, *et al.* Mining association rules using improved particle swarm optimization [J]. Surveying and Mapping Science, 2016, 41 (2): 168-172.

[12] Zhang X, Ding S, Sun T. Multi-class LSTMSVM based on optimal directed acyclic graph and shuffled frog leaping algorithm [J]. International Journal of Machine Learning & Cybernetics, 2016, 7 (2): 241-251.

[13] 杨旸，杨书略，柯闽. 加密云数据下基于 Simhash 的模糊排序搜索方案 [J]. 计算机学报，2017，40 (2)：431-444. (Yang, Yang Shulue, Ke Min. Fuzzy sorting search scheme based on Simhash under encrypted cloud data [J]. Acta Computer Science, 2017, 40 (2): 431-444.)

[14] Lee J, Kim H, Kim N R, *et al.* An approach for multi-label classification by directed acyclic graph with label correlation maximization [J]. Information Sciences, 2016, 351 (7): 101-114.

[15] Baxter J S H, Rajchl M, Mcleod A J, *et al.* Directed acyclic graph continuous max-flow image segmentation for unconstrained label orderings [J]. International Journal of Computer Vision, 2017, 123 (3): 415-434.